Volume 12 Number 1 1984 Nucleic Acids Research

Microcomputer programs for graphic analysis of nucleic acid and protein sequences

David W.Mount and Bruce Conrad

Departments of Molecular and Medical Microbiology and Biochemistry, College of Medicine, University of Arizona, Tucson, AZ 85724, USA

Received 15 August 1983

ABSTRACT

Four computer programs are described which allow two amino acid or DNA sequences to be compared for homology, the results being displayed in a 2-dimensional array on a printer page. The programs also may be used to visualize repeated sequences or dyad symmetry within a DNA sequence. Two of the four programs may be used with any printer, the other two require a printer with graphics capability. Many options are available including using only a portion of a sequence, specifying a window to demonstrate more significant structures and special restrictions on matching such as excluding the third base in a codon. Written in the C programming language, the programs run under the CP/M 80 operating system, and may be copied in binary format through a modem. They are also available for the IBM/PC.

INTRODUCTION

While a number of programs that automatically align protein or nucleic acid sequences or find internal structure within a molecule have been described (1-7), there has been somewhat limited application of these for microcomputers. Moreover, these same programs align according to a given set of criteria and may miss significant structural features, and are often limited in the length of sequence they can analyze. Another method uses a graphic matrix display in which two sequences are displayed on a printer page, one horizontally and one vertically, and structural features appear as diagonal lines on the page. This procedure, originally used for proteins, has been adapted for nucleic acids, and may also be used to show direct repeats and dyad symmetry within a DNA sequence (8). One set of programs written in Hewlett-Packard basic have been described (8). Here, we describe a similar set of programs which run under the standard CP/M operating system. In addition, the original source code, written in the highly portable C programming language, can be readily implemented on a variety of other computer systems.

DESCRIPTION OF PROGRAMS

If two nucleic acid or protein sequences share homology or if either has internal structures such as direct repeats of sequences or dyad symmetry, the property can be displayed by a printer in a two-dimensional array. We describe four programs which allow this type of analysis; baseplot, seqhom, protplot, and prothom.

System Requirements

Of the four programs we describe, baseplot and protplot print the nucleic acid base or amino acid letter, and should work with any printer. Seqhom and prothom provide a dot matrix analysis of two sequences, and require a printer with a graphics mode. Built into baseplot and protplot are instructions which make them usable with any printer. Seqhom and prothom require knowledge of how to make the printer enter and leave the graphics mode. They come configured to drive a NEC spinwriter 5510 or 5515 printer but we will describe on request how they may be made to work with other printers with a graphics mode.

Availability of programs

These programs are written in the C programming language, and have been compiled to run under the CP/M 80 operating system. They should run on any Z80 or 8080 computer with CP/M, including an Apple with the Z80 card. They have been downloaded into a binary format for transmission through the telephone. Transfer of this program form alleviates the need for a compiler. We will make this binary form available upon request. It may be copied through the telephone from a microcomputer in our laboratory with an answering modem, or if there are a sufficient number of requests, we will place these binary files on a commercially available host which can be reached through communication networks. These binary files (or hex files as they are called in the CP/M system) may be restored to an executable command form using the CP/M load command. Further adaptation is possible using the CP/M utility DDT. We have similar binary files that will run on the IBM/PC.

Sequence format

The DNA or amino acid sequences may be in a relatively free format:

1. any line with a semicolon in the first column is ignored.
2. spaces and tabs may be placed anywhere in the sequence
3. lines may be any length
4. sequences up to 8000 long can be analyzed.
5. for DNA sequences only A,C,G and T (capitals) are recognized. Lines with any other character are ignored.

6. for protein sequences only the standard one letter codes are recognized, plus Z for a termination codon. Lines with any other character are ignored.

7. a sequence range may be specified.

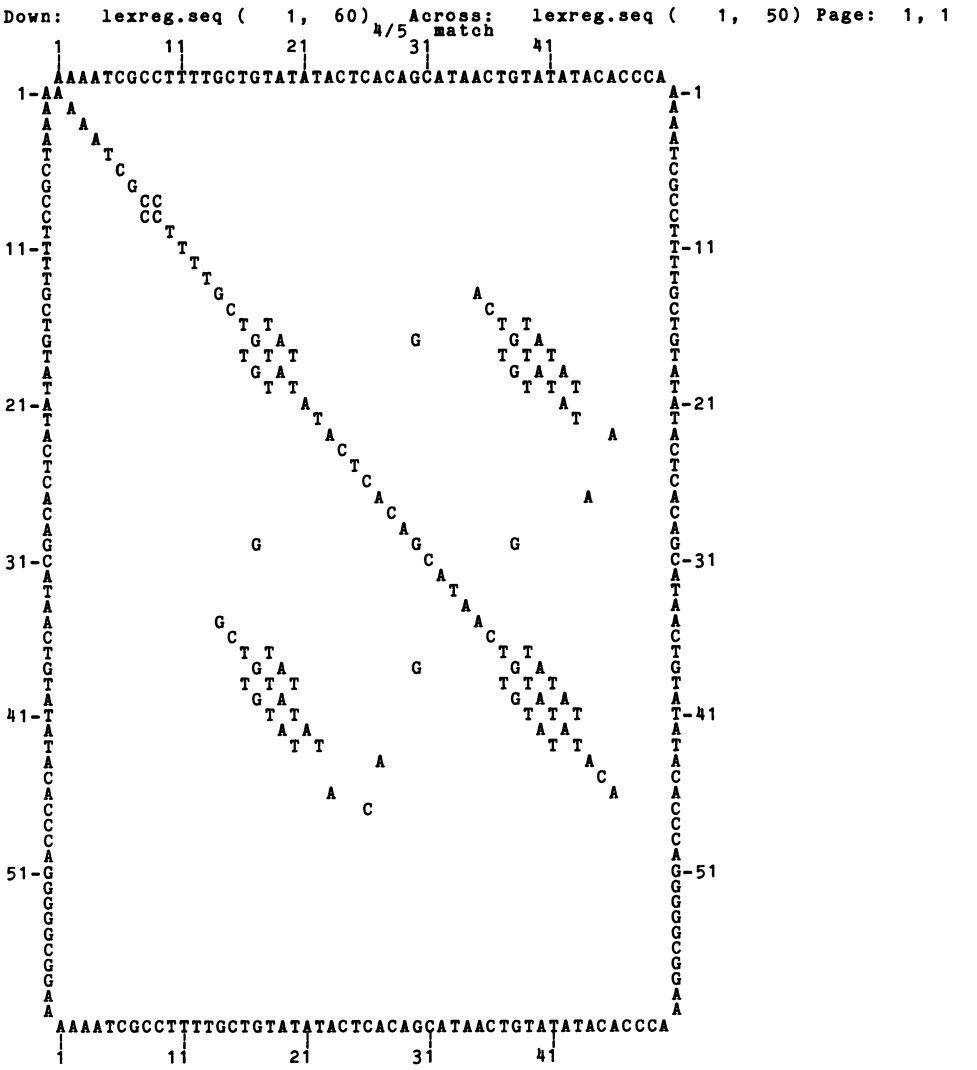

Figure 1.
Demonstration of direct repeat in DNA sequence with program baseplot. The sequence contains two tandem operators which are homologous in a natural sequence.

Baseplot

Baseplot displays a numbered portion of one sequence across the top and bottom of the page and a numbered portion of the other down the sides of the page. Where two bases are the same, the base from the horizontally displayed sequence is printed where the row and column containing those bases intersect. A window matching option which restricts printing of bases to only those which are at the 5' end of a given extent of homology. This option reduces the background printing of random matches when the two sequences are compared one base at a time and enhances printing of the more significant structures. A window size of up to 15 can be selected, along with a required number of bases to be matched within the window, and the option of not including every third base. The printing of a diagonal row of bases indicates homology.

A sequence may also be compared to itself to reveal the presence of direct repeats or to its complement to visualize dyad symmetry. In the last case, since the program is moving in opposite directions down the same molecule, symmetry should only be visible for one base at a time

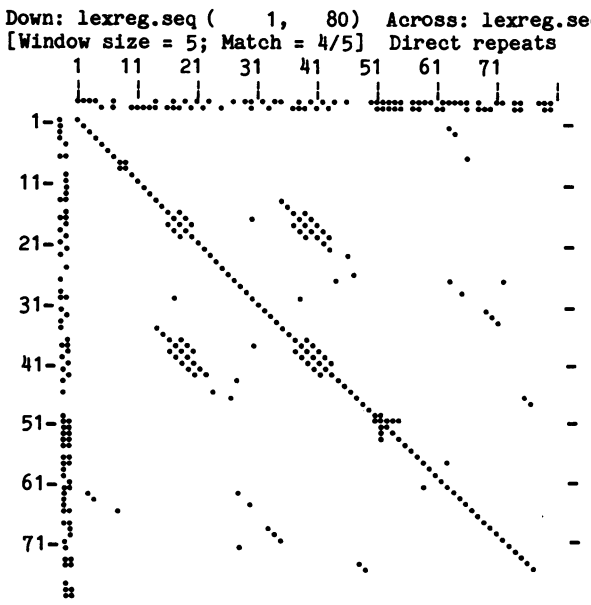

Figure 2.
Demonstration of a direct repeat in DNA sequence with an option of the program seqhom - same sequence as in Figure 1.

comparison, and should appear as upper right to lower left diagonals. To use the window option for observing symmetry with baseplot, it is necessary to create a reversed complement of one of the sequences. We can provide a program which performs these conversions on sequences. (Note that seqhom described below performs this reversal automatically by moving to the far end of the window and making its comparisons while moving backwards). Symmetries then appear as upper left to lower right diagonals.

In the above cases, the program moves sequentially through the sequences for the range of bases specified by the user, automatically advancing to the next page as each one is filled. An example of the use of baseplot for showing direct repeats within a sequence is shown in Figure 1.

Seqhom

The more advanced program (seqhom) uses a graphics mode to display more of the sequences on the page, prints the sequence along the edges of the page in a simple binary code (where 2 spaces are C, two periods G, an upper period A and a lower period T) and prints matches as periods or spaces. Like baseplot, it has a window feature and that improves

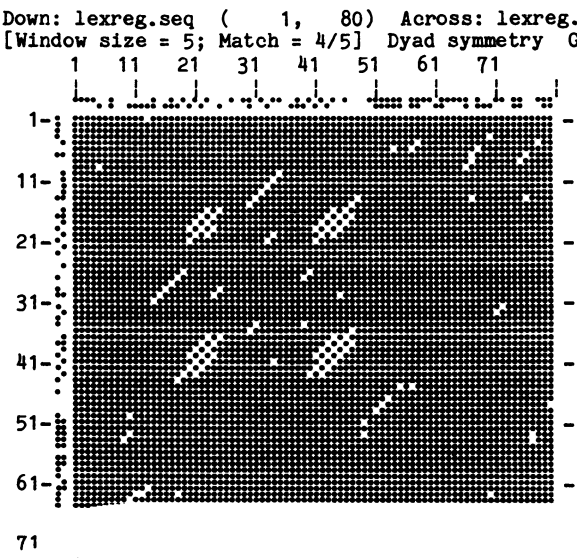

Figure 3.
Demonstration of dyad symmetry in a natural operator sequence with a second option of seqhom,same sequence as in Figure 1.

representation of the significant structural features. This more advanced program automatically moves backward or forward in the specified window within the sequence(s) to reveal homology, base pairing or symmetry; it is not necessary to generate a complementary sequence for the analysis. For a window size of 1, and searching for dyad symmetry, G:C base pair matches are shown by a comma and A:T pairs by a period. G:U(T) pairing can be optionally included. These features allow great flexibility in homology or symmetry analysis, but for dyad symmetry and secondary structure analysis are only useful for simple structures such as uncomplicated hairpin loops. Examples of the use of seqhom for finding direct repeats and dyad symmetry within a DNA sequence are shown in Figures 2 and 3, respectively.

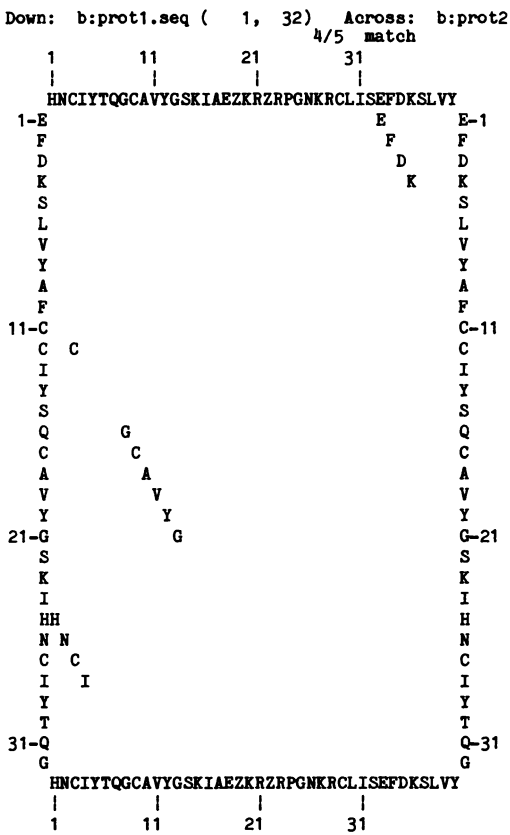

Figure 4.
Demonstration of repeat in a protein sequence. This sequence is a computer translation of a DNA sequence (see accompanying paper by Mount and Conrad).

Protplot

Protplot prints out the amino acids of two entered sequences on the printer page in their single letter code, one sequence across the page and the other down the side of the page and shows diagonals where there are matches or homologies. The output is similar to that of the baseplot program. There is also a similar window and match option, and the amino acid on the left in the matched region will be shown on the plot if the match condition is met. Only about 70X80 amino acids are shown on each page as opposed to about 180X240 with prothom (described below), but the advantage here is that the sequence can be displayed and read from the edge of the page. As with baseplot and seqhom, page advancing is automatic. An example of the output of the program protplot is shown in Figure 4.

Prothom

Prothom does the same thing as protplot except that it can throw a NEC spinwriter into and out of graphics mode to display homology as a dot plot. One again looks for upper-left to lower-right diagonals which represent homologies. As with seqhom, we can assist other users in adapting this program for other printers. We do not show an example of prothom output because it is very similar to that shown in Figure 2 for a DNA sequence.

ACKNOWLEDGEMENTS

This laboratory is supported by grants from the National Science Foundation and the National Institutes of Health.

REFERENCES

1. Korn, L.J., Queen, C.L., and Wegman, M.N. (1977) Proc. Natl. Acad. Sci. USA 74, 4401-4405.
2. Gingeras, T.K. and Roberts, R.J. (1980) Science 209, 1322-1328.
3. Sellers, P.H. (1979) Proc. Natl. Sci. USA 76, 3041-3045.
4. Smith, T.F. and Waterman, M.S. (1981) J. Mol. Biol. 147, 195-197
5. Goad, W.B. and Kanehisa, M.I. (1982) Nucl. Acids Res. 10, 247-263
6. Lipman, D.J. and Wilbur, W.J. (1983) Proc. Natl. Acad. Sci. USA 80, 726-730.
7. Zuker, M. and Stiegler, P. (1981) Nucl. Acids Res. 9, 133-148
8. Maizel, J.V., Jr., and Lenk, R.P. (1981) Proc. Natl. Sci. Sci. USA 78, 7665-7669